CUSTOMER PROFILING PADA SUPERMARKET MENGGUNAKAN ALGORITMA K-MEANS DALAM MEMILIH PRODUK BERDASARKAN SELERA KONSUMEN DENGAN DAYA BELI MAKSIMUM

Feri Sulianta
Universitas Widyatama
Jl.Cikutra No. 204A
feri.sulianta@widyatama.ac.id

Abstrak

Sebuah supermarket dengan sistem informasi berbasiskan komputer memiliki data transaksi dan data master yang di kelola dengan baik. Manajemen ingin membuat strategi bisnis berdasarkan penambangan historis data transaksi yang dimilikinya. Salah satunya dengan mencari tahu segmentasi pola perilaku pelanggan atau *customer profiling* yang memiliki daya beli tinggi terhadap barang-barang tertentu.

Untuk itu akan dilakukan pencarian informasi berharga terhadap analisa data mining dengan aturan klasterisasi menggunakan algoritma K-Means dalam mengelompokan pola belanja konsumen terhadap barang.

Kata kunci :

Data Mining, K-Means, Clustering, Customer Profiling

1. PENDAHULUAN

Data transaksi yang tersimpan dalam sistem basis data hendak dianalisa dalam menyingkapkan informasi berharga yang dapat digunakan sebagai strategi bisnis. Dalam hal ini perusahaan ingin melakukan profiling selera konsumen dalam membeli barang-barang tertentu berdasarkan besarnya daya beli konsumen.

Asumsinya, dengan tingginya daya beli konsumen terhadap barang tertentu maka ketersediaan barang tersebut akan meningkatkan pula omzet perusahaan, dan strategi dalam memberikan diskon terhadap barang tersebut dapat mengikat konsumen untuk tetap loyal terhadap perusahaan atau supermarket.

Penambangan data akan dilakukan dengan aturan klasterisasi menggunakan algoritma K-Means dan memilih sejumlah klaster yang akan memilah data pembelian barang kedalam kelompok tertentu berdasarkan karakteristik konsumennya.

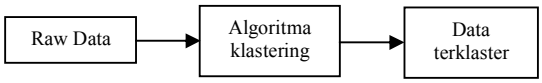
2. DATA, INFORMASI DAN KNOWLEDGE

Klasterisasi merupakan Knowledge Discovery Database (KDD) pada data mining yang akan membagi data kedalam kelompok-kelompok dimana pengelompokan yang terbentuk didasari pada obyek yang memiliki kesamaan karakteristik, sedemikian sehingga masing-masing elemen grup memiliki pembeda dengan grup lain.

Dengan alasan demikian, cara yang ditempuh untuk menyingkapkan temuan berharga dari data tersebut adalah dengan menggunakan data mining dengan aturan klasterisasi. Aturan klasterisasi dianggap cocok dalam kasus ini karena kemampuannya dalam memecahkan masalah penggolongan.

Obyek klaster yang ditinjau menjadi penentu bagaimana suatu item didistribusikan dalam kelompok tertentu dengan memperhitungkan derajat korelasi antara anggota klaster yang sama sehingga membentuk kumpulan data atau klaster.

Klaster akan mengungkapkan hubungan dan struktur di dalam data yang sebelumnya bias dan tersembunyi. Tetapi sewaktu ditemukan korelasi item-item dalam klaster, hasilnya sangat relevan dengan mempertimbangkan karakteristik data-data yang terklaster.



Gambar 1. Proses pengelompokan data secara umum kedalam klaster

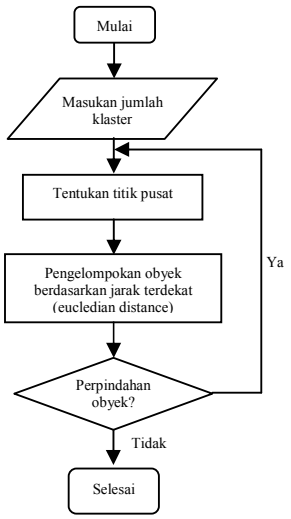
Tidak ada aturan spesifik data mining yang dipastikan akan menemukan data yang berharga, karena setiap aturan data mining sifatnya unik satu dengan yang lainnya. Tidak ada yang lebih baik dan lebih buruk, karena temuannya bisa berbeda satu dengan lainnya dan sama-sama berguna dalam membentuk pola data.

Proses pengelompokan data secara umum kedalam klaster tertuang dalam algoritma klastering, dalam hal ini digunakan algoritma K-means.

Algoritma k-means mendasari metode operasinya berdasarkan namanya yaitu ‘mean’ atau rerata. Dimana akan dilakukan pengamatan ke dalam kelompok k, dimana k diberikan sebagai parameter masukan. Kemudian dilakukan pengamatan pola untuk setiap kelompok berdasarkan kedekatan pada rerata jarak klaster yang adalah titik tengah, selanjutnya titik pusat atau *centroid* akan dikalkulasi ulang dan proses dimulai kembali.

K-Means adalah teknik yang dikategorikan *greedy*, komputasi dilakukan dengan teknik yang efisien, menjadi salah satu

algoritma klastering yang paling banyak digunakan untuk klasterisasi [7][8].



Gambar 2. Flowchart Algoritma K-Means

Algoritma K-Means dapat dijabarkan sebagai berikut [1][6][8]:

1. Algoritma secara acak akan memilih beberapa k titik sebagai pusat klaster.
2. Setiap titik dalam dataset didedikasikan untuk klaster tertutup yang didasarkan pada jarak *Euclidean* antara setiap titik dan masing-masing pusat klaster.
3. Setiap pusat klaster dikalkulasi ulang sebagai rerata dari titik-titik pada klaster.
4. Langkah 2 dan 3 diulangi hingga klaster berkumpul dan tidak diapati perubahan ketika langkah 2 dan 3 diulang atau bahwa perubahan tidak membuat perbedaan karakteristik nilai data pada member klaster. Pada kondisi ini klaster mencapai taraf stabil, dalam pengertian tidak ada lagi obyek yang dapat dipindahkan.

Jarak *Euclidean* antara dua titik atau obyek atau item x dan y, dimana jika x = (x1, x2,..., xn) dan y = (y1, y2,..., yn) adalah dua titik pada ruang lingkup *Euclidean* n-area, maka jarak dari x ke y, atau dari y ke x ditulis dalam formula sebagai berikut:

$$d(X,Y)=\sqrt{(|Y_1-X_1|^2+|Y_2-X_2|^2+...+|Y_n-X_n|^2)}$$

Keterangan:

- Y: *Centroid* yang menjadi pusat cluster
- n: Jumlah dimensi dari pola masukan

Formula kalkulasi titik pusat dengan mencari rata-rata klaster:

$$R_k = \frac{1}{N_k}(X_{1k}+X_{2k}+...+X_{nk}).$$

Keterangan:

- Rk: rata-rata baru
- Nk: jumlah training pattern pada cluster k
- Xnk: pola ke –n yang menjadi bagian dari klaster k

Pada algoritma klaterisasi, fokus ditujukan pada penentuan jumlah klaster atau k. Jumlah klaster yang ingin dibentuk ini akan digunakan sebagai masukan bagi algoritma. Pada dasarnya algoritma tidak mampu menentukan jumlah klaster dan ini bergantung sepenuhnya pada pengguna untuk mengidentifikasi terlebih dahulu jumlah klaster yang diinginkan[3][4].

Tidak mudah menentukan banyak klaster, dan ini adalah strategi yang dipilih dengan asumsi bahkan pertimbangan yang sifatnya intuitif.

Misalnya, jika kita memiliki sejumlah data orang yang teralamati sebagai orang dengan status ‘kawin’ dan ‘lanjang’. Jika menentukan algoritma k-berarti dengan k=2, hal ini akan secara tegas terklaster, tapi jika k=3, maka kita akan memaksa orang untuk dialokasikan ke dalam tiga kelompok. Dan jika memilih k=4 atau lebih, maka relevansinya tidak teralamati jelas, karena memaksa obyek untuk dialokasikan dalam empat klaster, meskipun demikian hal ini dapat dilakukan dengan asumsi adanya atribut-atribut lain yang dijadikan parameter untuk memisahkan kelompok orang.

Atas dasar inilah, eksperimen dilakukan untuk memilih beberapa nilai k guna mengidentifikasi nilai yang paling sesuai untuk mengelompokan obyek.

3. IMPLEMENTASI

Berdasarkan data transaksi supermarket yang akan dianalisa untuk keperluan mining, berikut ini langkah-langkah yang akan dilalui guna mencapai tujuan akhir yaitu melakukan klasterisasi terhadap sejumlah barang berdasarkan daya beli konsumen.

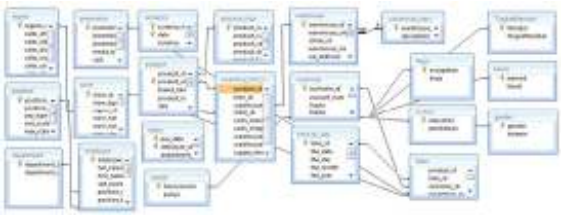
Akan dipilih barang yang menjadi pilihan konsumen yang memiliki daya beli tinggi, dimana supermarket akan menyediakan barang-barang tersebut lebih banyak dan menerapkan strategi bisnis terhadap barang-barang tertentu untuk mengikat konsumen tetap loyal yang tujuan akhir adalah peningkatan omzet.

Mendapatkan data merupakan langkah pertama yang akan dilakukan adalah melengkapi kebutuhan analisa data, yakni dengan memilih data transaksi dan data master.

Guna mendapatkan kelengkapan data, maka akan dilakukan peninjauan basis data. Pada dasarnya data yang diperoleh tidak didapati kesalahan data, tidak pula didapati *missing value* karena data sudah terpelihara dengan baik dalam sistem basis data dengan penyangga validitas basis data yakni *Referensial Integrity Rule.*

Beberapa atribut pada tabel konsumen perlu dikembangkan lebih lanjut guna memetakan data nominal kedalam data numerik sehingga menjadi data yang dapat diukur pada proses klasterisasi misalnya: tabel rumah merupakan pemetaan nilai Y dan N, dimana Y dikonversi menjadi 2 dan N dikonversi menjadi 1 dengan asumsi orang yang memiliki rumah pasti berada dalam taraf kemakmuran dan memiliki daya beli tinggi.

Hal ini dilakukan dengan menambahkan informasi antara lain: Tingkat Member, Kawin, Kerja, School dan Rumah. Kelengkapan tabel yang dibutuhkan terlihat dalam *Database Environment* sebagai berikut di bawah:



Gambar 3. Skema basis data Foodmart.mdb

Berikut ini beberapa informasi perihal konsumen yang telah dipetakan kedalam nilai numerik, antara lain:

- Jumlah barang yang dibeli dalam satu kali transaksi.
- Jumlah anak, yakni semakin banyak anak maka tingkat konsumsi semakin tinggi.
- Status kawin dipertimbangkan dengan justifikasi bahwa orang-orang yang berstatus kawin memiliki tanggung jawab lebih untuk memenuhi kebutuhan keluarga. Nilai atribut yang dipetakan adalah:
    - Tidak Kawin : (1)
    - Kawin : (2)
- Jenjang profesi dijadikan pertimbangan bahwa semakin tinggi profesi, daya beli semakin besar. Jenjang profesi memiliki nilai atribut yakni:
    - Manual: (1)
    - Skill Manual:(2)
    - Managemet: (3)
    - Professional: (4)
- Pendapatan, yakni semakin besar pendapatan, semakin besar daya beli. Atribut pendapatan dikonversi dari rentang harga nominal menjadi besaran numerik:
    - $10K-$30K: (30)
    - $30K-$50K: (50)
    - $50K-$70K: (70)
    - $70K-$90K: (90)
    - $90K-$110K: (110)
    - $110K-$130K: (130)
    - $130K-$150K: (150)
    - $150K+: (200)
- Kepemilikan rumah. Kepemilikan rumah menjelaskan derajat kemakmuran yang diukur dengan jumlah rumah yang dimiliki. Dengan nilai atributnya yakni:
    - Y: (2)
    - N: (1)
- Jumlah kepemilikan mobil. Justifikasinya yakni kepemilikan mobil berbanding lurus dengan kemakmuran dan daya beli.
- Pendidikan. Semakin tinggi pendidikan maka semakin tinggi derajat kermakmuran dan daya beli pun tinggi, nilai atribut yakni:
    - Graduate Degree: (5)
    - Bachellors Degree: (4)
    - Partial College: (3)
    - High School Degree: (2)
    - Partial High School: (1)
- Tingkat Member. Semakin tinggi level keaggotaan, maka tingkat daya beli pun semakin tinggi, dengan nilai atribut yang digunakan yakni:
    - Gold: (4)
    - Silver: (3
    - Bronze: (2)
    - Normal: (1)

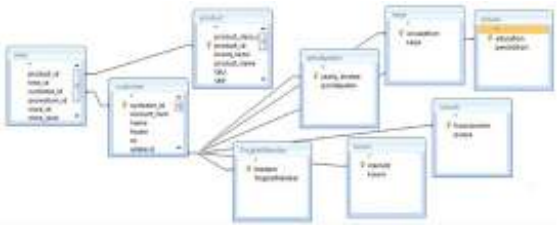
Berdasarkan atribut bentukan diatas, didapat 10 Attribut yang digunakan untuk mengalamati pola belanja konsumen terhadap barang, yaitu: *product_name, unit_sales, total_children,* kerja (derajat profesionalitas pekerjaan), pendapatan, *num_cars_owned*, pendidikan, punya (kepemilikan rumah), kawin (status kawin atau tidak), dan tingkat member.

Setelah kelengkapan data yang hendak dianalisa didapat, maka akan dilakukan ekstraksi data dengan melakukan query dan konversi data kedalam format CSV atau Comma Separated Value yang dapat diterima Tool Data Mining Weka.

Perintah SQL dalam melakukan Query data:

```
SELECT sales.time_id,
product.product_name, sales.unit_sales,
customer.total_children, kerja.kerja,
pendapatan.pendapatan,
customer.num_cars_owned,
school.pendidikan, rumah.punya,
kawin.kawin FROM TingkatMember INNER JOIN
(kawin INNER JOIN (school INNER JOIN
(rumah INNER JOIN (kerja INNER JOIN
((customer INNER JOIN (sales INNER JOIN
product ON sales.product_id =
product.product_id) ON
customer.customer_id = sales.customer_id)
INNER JOIN pendapatan ON
customer.yearly_income =
pendapatan.yearly_income) ON
kerja.occupation = customer.occupation)
ON rumah.houseowner =
customer.houseowner) ON school.education
= customer.education) ON kawin.married =
customer.marital_status) ON
TingkatMember.Member =
customer.member_card;
```

Dalam proses query akan dihasilkan flat file yang menjadi dasar kelengkapan data yang akan dianalisa. Jumlah datanya tercatat sebanyak 65535 *instances*.



Gambar 4. Tabel basis data yang digunakan dalam query data

Data query ini akan ditranslasikan dalam file dengan format CSV yang kemudian akan dilanjutkan dengan mengakses data tersebut menggunakan *WEKA Data Mining Tool* yang selanjutnya dilakukan proses klasterisasi.

Tool Data Mining WEKA digunakan dalam membaca data sebelum melakukan pemrosesan klasterisasi. Dalam hal ini dipilih 5 klaster guna mendapatkan pengelompokan +/- 20% dari data transkasi, yang sebelumnya dicoba 10 klaster yang hanya menghasilkan pengelompokan dengan +/- 9 % data saja per-klasternya. Status keanggotaan pun menjadi salah satu pertimbangan, dimana konsumen digolongkan dalam 4 keanggotaan. Pembagian kedalam 10 klaster akan membagi data terlalu kecil sedangkan pengklasteran 5 relevan dengan komposisi data yang didapat.

Tabel 1. Komposisi instace dalam klaster

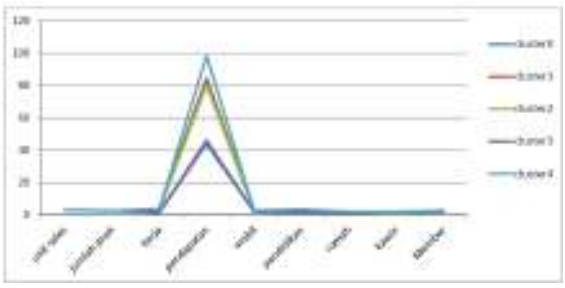
No	Klaster	Banyaknya instance	Persentasi instance dalam klaster (%)
1	0	16197	25%
2	1	10905	17%
3	2	13713	21%
4	3	12985	20%
5	4	11735	18%

Selanjutnya proses pengklasterisasi dimulai setelah memilih algoritma dan jumlah cluster yang diinginkan. Pada *Weka Data Mining Tool* , dipilih algoritma klaster: Simple K-Means dengan jumlah klaster: 5.

Data hasil klasterisasi *Data Mining Tool* ini dianalisa lebih lanjut dengan mentranslasikan data pada Microsoft Excel, dimana setiap cluster yang terbentuk dianalisa perbedaan karakteristiknya dengan klaster lainnya dalam implementasi grafis, acuan yang digunakan adalah data hasil rerata dari data mining tool.

Tabel 2. Instance terklaster

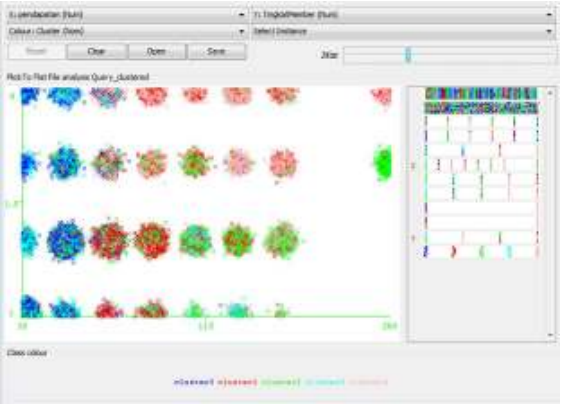
area	unit sales	jumlah anak	kerja	pendapatan	mobil	pendidikan	rumah	kawin	member
c0	3.109588195	2.620485275	1.535778	43.7408162	1.826326	1.554979	1.000556	1.302525	1.822991912
c1	3.100137552	2,558,367,721	3.735122	84.26776708	2.281522	2.621183	1.000183	1.352774	2.242090784
c2	3.07306935	2.40058339	2.697222	80.2165828	2.148326	2.410413	1.999927	1.000146	1.967403194
c3	3.11728918	2.588602233	1.626646	46.28648441	1.976588	1.540624	1.999923	1.999846	1.988294186
c4	3.093225394	2.585087346	3.895611	98.30847891	2.632382	3.394291	1.999915	1.999318	2.710779719



Gambar 5. Diagram karaktersik lima klaster berdasarkan sembilan atribut

Keterangan :

1. Cluster_0: orang-orang yang membeli produk pada klaster ini memiliki krakteristik dan daya beli ekonomi rendah, ini teralamati dengan tingkat pendapatan yang rendah, tingkat member dan kepemilikan kendaraan yang rendah, tetapi memiliki jumlah anak yang tinggi dibandingkan dengan klaster lainnya.
2. Cluster_1: kelompok ini memiliki karakteristik konsumen dengan tingkat pendapatan tinggi meskipun tidak setinggi kelompok Cluster_4. Tingkat pendidikan tertinggi konsumen berada pada klaster ini.
3. Cluster_2: memiliki karakteristik barang yang dibeli oleh konsumen dengan daya beli menengah, tingkat pendapatan dan keanggotaan yang berada di level yang mapan, meskipun tidak tergolong pada kriteria menengah keatas. Cluster_2 memiliki karakteristik dengan daya beli sedikit lebih rendah dibandingkan Cluster_1. Pembeda yang terlihat jelas adalah tingkat kepemilikian rumah yang lebih tinggi dibandingkan Cluster_1.
4. Cluster_3: klaster ini memiliki karakteristik konsumen berdaya beli sedikit lebih tinggi dengan kelompok Cluster_0, dengan jumlah kepemilikan anak lebih rendah dibanding Cluster_0.
5. Cluster_4: pada klaster ini, karakteristik konsumen berada pada tingkat daya beli ekonomi yang tinggi, ini diperlihatkan korelasi kuat antara tingginya tingkat pendapatan dan tingkat keanggotaan, tingkat pekerjaan dan kepemilikan mobil, sedangkan atibut lain menempati level yang tinggi ,meskipun bukan yang tertinggi. Salah satu contohnya diperlihatkan pada gambar di bawah yang memperlihatkan korelasi dua atribut member dan pendapatan.



Gambar 6. *Cluster Visual Inspection* - korelasi antara pendapatan dan tingkat member yang terklaster pada cluster_4 yang adalah klaster ke-5

Setelah dianalisa, maka dapat dirumuskan bahwa perencanaan strategi bisnis ditujukan untuk klaster ke-lima (Cluster_4) yang berupa kelompok barang-barang yang dibeli oleh konsumen dengan kemampuan atau daya beli tinggi yang tercermin dari tingkat keanggotaan dan juga pendapatan diatas rata-rata disamping jenis kepemilikan lain yang dengan nilai yang tinggi seperti kepemilikan perumahan, tingkat pendidikan, tingkat kebutuhan berdasarkan status kawin dan jumlah anak.

Sorting dan pengelompokan nama produk pada klaster ke-5 (Cluster_4). Untuk melihat barang-barang yang terklaster pada kelompok ini, dilakukan *sorting* berdasarkan nama produk dan dikelompokan untuk didapat total jumlah barang-barang yang dibeli untuk masing-masing produk, selanjutnya diurut untuk mendapatkan ranking barang paling banyak dibeli yang dijadikan prioritas sebagai produk yang diminati dan sesuai dengan selera konsumen berdaya beli tinggi.

4. TUJUAN DAN KESIMPULAN

Dengan demikian hasil klasterisasi yang membagi data kedalam 5 bagian didapatkan karakteristik konsumen terhadap data yang dibeli dimana semua data yang terklaster pada klaster ke-5 memiliki porsi 18% dari keseluruhan data transaksi dengan 11735 *instances* yang ditinjau, merupakan barang-barang yang dibeli oleh konsumen yang memiliki daya beli tinggi, hal ini teralamati dengan tingginya tingkat pendapatan dan keanggotaan pada kelompok konsumen ini, disamping atribut lain pun memiliki tingkat yang relatif tinggi meskipun tidak mendominasi.

Barang yang tergolong klaster lima adalah barang yang diminati oleh konsumen yang memiliki daya beli tinggi, dengan demikian menambah pengadaan barang untuk kategori barang ini akan meningkatkan omzet perusahaan. Strategi bisnis dapat dilakukan untuk mengikat konsumen pada kategori ini, seperti pemberikan diskon atau sejenisnya.

TOP 15 item barang yang menjadi pilihan konsumen berdaya beli tinggi dapat dilihat pada Tabel 2.

Tabel 3. Daftar barang yang diminati konsumen berdaya beli tinggi

Nomor	Nama Produk
1	'Urban Egg Substitute'
2	'Atomic White Chocolate Bar'
3	'Great English Muffins'
4	'Robust Monthly Home Magazine'
5	'Denny AAA Size Batteries'
6	'Big Time Frozen Chicken Thighs'
7	'Carlson Havarti Cheese'
8	'Excellent Orange Juice'
9	'Faux Products Deodorant'
10	'Hermanos Beets'
11	'Super Chunky Peanut Butter'
12	'Nationeel Golden Raisins'
13	'Red Wing Room Freshener'
14	'High Top Elephant Garlic'
15	'Big Time Beef TV Dinner'

5. DAFTAR PUSTAKA

[1] D T Pham , S S Dimov, and C D Nguyen. Selection of K in K-means clusteringProc. IMechE Vol. 219 Part C: J. Mechanical Engineering Science, Cardiff University, Cardiff, UK. 2004.

[2] Dokumentasi Machine Learning Weka Tool. Help Tool WEKA. 2012

[3] Farajian, Mohammad Ali., Mohammadi, Shahriar. Mining the Banking Customer Behaviour Using Clustering and Association Rules Methods. International Journal of Industrial Engineering and Production Research.Vol 21, Number 4 pp. 239-245.2010.

[4] Ronald B. Larson. New Market Groupings Based on Food Consumption Patterns, Department of Marketing, Haworth College of Business, Western Michigan University. Agribusiness, Vol. 20 (4) 417–432 Wiley Periodicals, Inc.Published online in Wiley InterScience. 2004

[5] Sharma, Narendra. Comparison the various clustering algorithms of weka Tools, International Journal of Emerging Technology and Advanced Engineering. Volume 2, Issue 5, May 2012.

[6] Sri Andayani. Pembentukan cluster dalam Knowledge Discovery in Database dengan Algoritma K - Means. Jurusan Pendidikan Matematika FMIPA UNY.

[7] Techniques of Cluster Algorithms in Data Mining. Data Mining and Knowledge Discovery, 6, 303–360, 2002 Kluwer Academic Publishers. 2002.

[8] Witten,Ian., Frank, Eibe. Data Mining – Practical Machine Learning Tool and techniques. Morgan Kaufmann Publishers. 2005